Evaluating Lossiness and Fidelity in Information Visualization

Richard Brath*[a], Ebad Banissi[a]
[a]London South Bank University, 103 Borough Road,
London SE1 0AA, UK

ABSTRACT

We describe an approach to measure visualization fidelity for encoding of data to visual attributes based on the number of unique levels that can be perceived; and a summarization across multiple attributes to compare relative lossiness across visualization alternatives. These metrics can be assessed at design time in order to compare the lossiness of different visualizations to aid in the selection between design alternatives. Examples are provided showing the application of these metrics to two different visualization design scenarios. Limitations and dependencies are noted along with recommendations for other metrics that can be used in conjunction with fidelity and lossiness to gauge effectiveness at design-time.

Keywords: Information Lossiness, Visualization Fidelity, Visual Attribute Permutations.

1. INTRODUCTION

Information visualization (infovis) transforms data attributes into visual attributes. A bubble plot, for example, transforms on data attribute to the position of a glyph along the x-axis, a second data attribute is transformed to position along the y-axis, a third data attribute is transformed to dot size, and possibly a fourth data attribute is transformed to brightness. This is a fundamentally lossy process because visual attributes may encode information with lower number of perceivable levels than the original data [1]. For example, encoding data as brightness may result in perceptual estimation errors as much as 20% meaning that brightness should be limited to encoding only two to four numeric levels [2]. A means to measure the potential lossiness can be an effective design time evaluation tool for assessing alternative visualization designs. A novel approach for evaluating potential information lossiness is presented based on first evaluating the fidelity per visual attribute in an encoding; and then estimate permutations across multiple attributes to compare relative lossiness.

2. BACKGROUND

In an evaluation of infovis heuristics, Forsell and Johansson [3] identify Information Coding (originally defined [4]) as the most frequent heuristic for explaining usability problems; that is, a problem in the mapping of data elements to visual attributes, such as inappropriate encoding. Chen and Floridi [1] provide an analysis of information visualization from a philosophy of information perspective comparing a communication system with a visualization pipeline. Like communication, the visualization pipeline is subject to information loss, error and noise across each step in the pipeline.

MacKinlay [5] defined *expressiveness* as a visualization that encodes all relevant information and only that information. In MacKinlay's expressiveness, an encoding transforms a data attribute in a one-to-one correspondence without removing or adding data. MacKinlay's effectiveness can be encoded as a grammar; which has been expanded upon by future authors (e.g. Wilkinson, Heer & Bostock). MacKinlay defines effectiveness as a separate consideration, broadly meaning that the presentation is clear. Whereas MacKinlay's effectiveness is like a grammar, effectiveness is akin to semantics.

One approach to selecting between different visual attributes is to create a ranking of visual attributes by effectiveness. In MacKinlay's APT, effectiveness is defined as the assessment that some visual attributes will be more suited to some types of data than others, for example, MacKinlay presents a table ordering different visual attributes by effectiveness for different types of data (i.e. categoric, ordered, quantitative).

Visual attributes have different properties and effectiveness for different types of encoding and have been expressed by other researchers as well, e.g. [6,7,8,9]. The work of Bertin [6] characterized different visual attributes by *length* which indicates the number of unique levels that can be perceived for a given attribute. Bertin derived his recommendations for length based on his experience with printed visualizations, so for example, position of a mark along the plane is

considered by Bertin to support 10 perceptible levels per centimeter. Size variation provides up to 20 perceptible levels. For some visual attributes Bertin does not provide the number of levels when mapping quantitative data to the attribute; but does provide levels when considering the visual attribute for depicting categories, i.e. where discrete data categories must be differentiated. For Bertin, brightness (value) provides up to six levels, texture provides four to five levels, color is considered to have eight distinct hues and orientation has four levels. For Bertin, shape has an infinite number of levels, but does not offer any ability for association perception - i.e. objects of the same shape will not pop-out across a field but only associate with other shapes in the same vicinity.

Beyond Bertin and visual attribute lists, some research has measured accuracy of visual judgment for quantitative data for a few visual attributes [10,11], such as Cleveland and McGill, who experimentally established error rates for perceptual estimation of visual attributes such as line lengths

Rule-based systems for effectiveness have been used for automated visualizations. For example, AutoVisual [12] uses *potential effectiveness* (MacKinlay's effectiveness extended for interaction) and uses a priority ordering of variables in relation to task and mapping to either explicit representations (of the immediate inner world) or interactions such as the outer world or exploratory tools. AutoVisual optimizes for effective encoding, minimal required interaction, and fast response time; with additional rules to ensure legibility and also lowers priority ordering data variables with few levels (i.e. a data attribute with fewer than five unique values). VISTA [13] uses a set of rules to validate encodings. At a high level, composition rules determine combinations of multiple visual encodings; such as rules for merging marks, superimposition, union, transparency or intersection. At a lower level to assess effectiveness, the system has 150 rules for visual perception [14] e.g. quantitative data is better mapped to geometry than color.

User task, such as awareness, exploration or analysis, is an important consideration for effectiveness. For example, if the task requires rapid awareness among hundreds of indicators (e.g., hundreds of glyphs), a blinking indicator may be a highly effective encoding drawing immediate attention to it. For an analytical or exploratory task, blinking would be considered distracting whereas the ability to easily perceive differences in magnitude is important to these tasks. Amar and Stasko consider the analytic task to be more important the representational primacy [15].

For analytic tasks, such as the perception of differences in magnitude, visual fidelity is an important issue: some visual attributes only encode a few levels of differentiation. Lossiness occurs when the number of discrete data values that need to be shown is greater than the number of levels that can be perceived with the target visual attribute. Assessing the difference is the objective of the approach in this paper. This approach is an attempt to assess the visualization quality although this approach is at a scale that attempts to assess quality and tradeoff decisions across very different kinds of visualization encodings rather than optimizing a particular visualization type (e.g. [16]).

The four-level nested design model for design and validation for visualization by Munzner [17] provides levels stepping through 1) domain problem characterization, 2) data and task abstraction, 3) encoding/interaction techniques, and 4) algorithm design. The approach in this paper is specifically a measure for evaluation of the quality of visual encoding of data, that is, between-levels of data abstraction and visual encoding.

A broad survey of visualization quality metrics are presented in [18], particularly with regards to measuring patterns generated by visualizations for high-dimensional data with the goal of helping users view the best configurations. However, the approach in this paper is focused on the visual mapping stage of the visualization pipeline (not data transformation nor view transformation); and in particular, on design-time evaluation of lower-dimensional visualization design alternatives, each with different visual mapping configurations and/or potentially novel visualizations.

There are some similarities to [19] with metrics such as maximum number of dimensions and dimensional score; but those were measures of the complexity of the encoding, not measures of information lossiness.

The unique contribution of this approach is that it goes beyond visual attribute ranking or rule based approaches for design alternatives. Instead this new approach measures the visual fidelity of the attribute encoding per attribute and provides a summarization of permutations across attributes to estimate a comparative lossiness between visualization design alternatives.

3. FIDELITY AND LOSSINESS

The approach outlined in this paper builds incrementally by first evaluating each data dimension and corresponding *fidelity* of the visual encoding; and then combines these to evaluate multi-dimensional encodings to assess the comparative *lossiness* of different designs with potentially significantly different encodings.

In this paper the term *fidelity* denotes the number of unique levels perceivable for a particular visual attribute, in the context of the source data. For example, in a visualization that uses shape to differentiate between two categories (e.g. male, female mapped to two unique shapes, e.g. ♂, ♀), the shape attribute would have a fidelity level of two - accurately representing the two data levels.

Lossiness is comparative metric. It is based on computing the total permutations of fidelity per each visual attribute in a design. The total permutations per design are normalized to show the relative lossiness across design alternatives.

3.1 One Dimensional Fidelity

An information visualization encodes a number of different dimensions of data. A simple bar chart or pie chart encodes two dimensions of data: a set of categories and a set of values corresponding to those categories. In figure 1, bar chart shows values along a common scale making the difference between black currants and cherries quite visible, with the size difference on the pie chart may not be discernable.

Fruit	Qty
Apples	33
Black Currants	19
Cherries	17
Figs	13
Guava	11
Kiwis	7

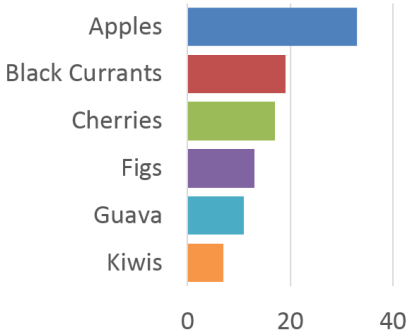

Figure 1. Pie vs. bar. Pie and bar each show the same data, each use the same area, but differences in similar sizes are more perceivable in the bar chart.

While these differences may be intuitively understood, the difference can also be articulated as a measurement of the number of discrete perceptible levels. With regards to quantities, Cleveland and McGill [10] and later Heer and Bostock [11] provide metrics for accuracy judgments and error rates for different visual attributes, such as an average error rate of ± 4.5% on angle judgments vs. an error of ± 2.5% on adjacent length judgments aligned to a common scale. As such, the pie chart may be considered as having fewer perceptible levels than the bar chart. With regards to quantities, the bar chart provides a higher fidelity of data encoding than the pie chart and therefore has lower information lossiness.

To show this quantitatively using Cleveland and McGill's error rates, in the case of the bar chart, the bar *figs* is 39% of the length of the longest bar and the bar *guava* is 33% of the length of the longest bar, beyond the ±2.5% error rate for length discrimination. The bar lengths provide a fidelity of six uniquely perceivable levels. In the case of the pie chart, the angle subtended by the category *figs* is 13% of the total whereas *guava* is 11% of the total - within the range of ± 4.5% error rate on angle judgments. Similarly, the angles for *black currants* and *cherries* are close enough to result in potential error estimation as well. As a result, the pie wedges provide a fidelity of four uniquely perceivable levels. The bar chart, with a length fidelity at six levels, is superior to the pie chart, with an angle fidelity of four levels.

Error rates have been established for only a few visual attributes. Using Bertin's levels instead, one can inspect the bar chart and establish that the difference between any pair of bars is more than one millimeter, implying all lengths are clearly distinguishable, i.e. having six uniquely perceivable levels. For orientation, Bertin considers discrete angles of 30° increments distinguishable. In the pie chart, the angles of the wedges are 25, 40, 47, 61, 68 and 118 degrees. The angles for *figs* and *guava* at 40° and 47° are very close together resulting in potential error estimation; whereas the angle for *kiwis* at 25° is 15° difference from the next smallest wedge, half of the 30° increment and potentially not subject to error. Following this approach, *black currants* and *cherries* are also close and subject to error, and overall only 4 levels

are perceivable for the pie chart. Using Bertin's values, the fidelity levels are the same as the results using Cleveland and McGill's error rates, that is, the bar chart lengths have six levels and the pie chart has four levels.

Lost information can be retrieved via interactions such as tooltips, however, tooltips are much slower than preattentive perception (the visual pop-out of lengths, areas and angles); and slower than simply shifting attention and reading a label already visible. This approach seeks to measure what is visible, not what is hidden and/or accessed via interactions.

3.2 Mutli-Dimensional Fidelity

Consider three design alternatives shown in figure 2. The end-user, a financial expert, needs a visual display of news headlines. Data of interest includes the news headline, recency and readership. It is important to note that the user community is interested in all the headlines - a headline with low readership may still be of interest to a particular client.

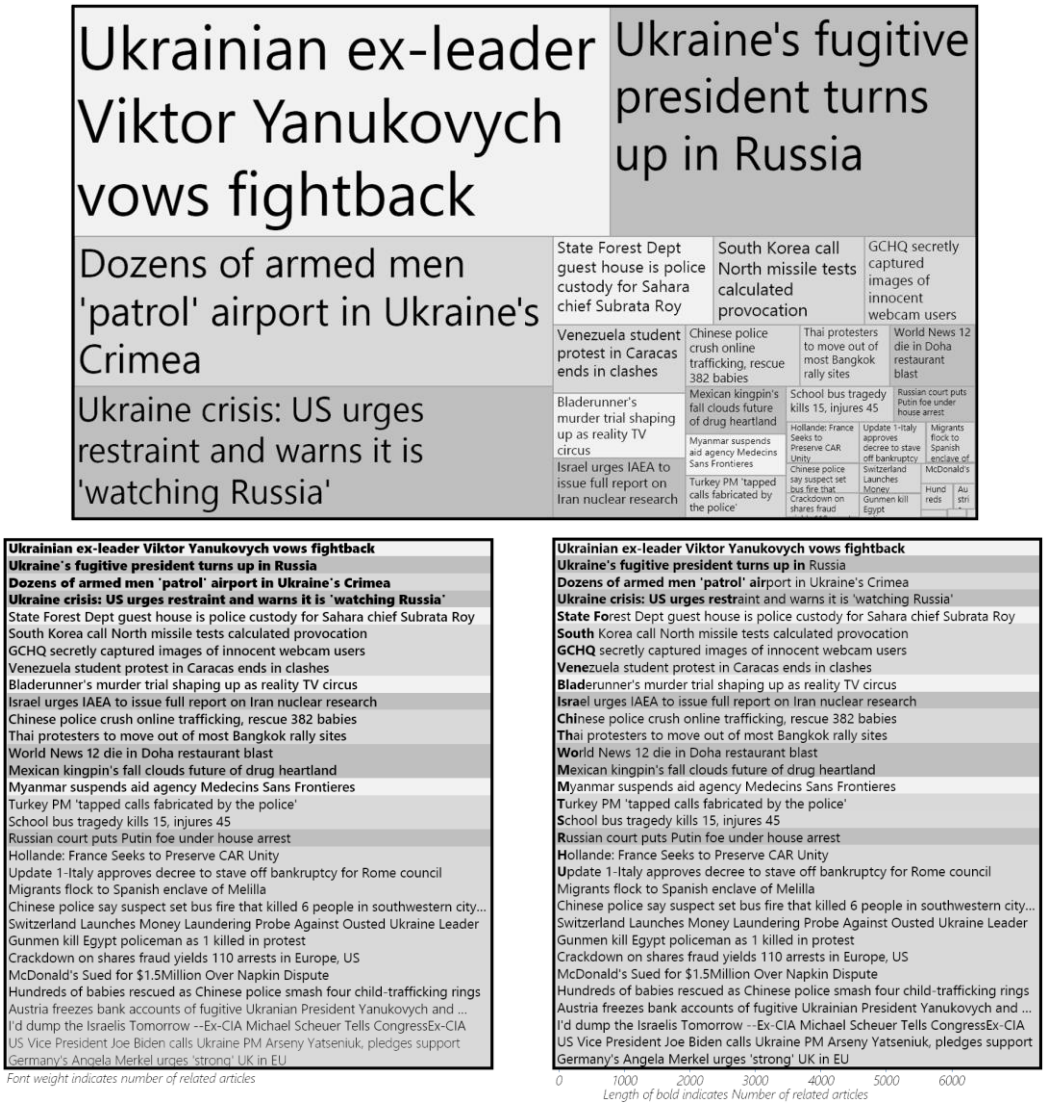

Figure 2. News headlines. Top: treemap, based on [20]. Size indicates readership, color indicates recency. Lower left: Same headlines and shading with font weight indicating readership. Lower right: Same headlines and shading with proportional length of bold indicating readership [21].

In all figures, the representation uses text to encode the literal headline and brightness of the background to encode recency. The only difference is the encoding of readership: figure 2 top uses a treemap sizing text (redrawn based on newsmap.jp [20]), figure 2 left uses font weight and figure 2 right uses proportional string encoding. Proportional string encoding applies a type attribute, such as bold, along a portion of a text string to indicate quantity with the proportion

shown along a common scale. The different encodings have implications in the number of levels perceivable, detailed below and summarized in table 1:

- *Recency:* is encoded as the brightness of the background behind each headline. The same brightness encoding is used in all three examples and always applied to the background area behind the headline. For the purposes of evaluation comparative designs, the number of levels is irrelevant as it will be the same across all three variants and have a neutral effect on the result. However, to estimate the number of levels of brightness, according to Ware [2], is two to four levels. Since the brightness range is constrained (i.e. not too dark) in order to keep the text on top of the background readable, the number of levels of brightness may be considered to be slightly lower, at three distinct levels.
- *Readership:* is encoded in the treemap as area. Based on visual inspection and Heer's error rates (6% error rate [11]), there should be on the order of 10-14 uniquely perceivable sizes in the treemap. In the font weight example, the chosen font has 5 distinct weights. Font weights have not been formally evaluated, e.g. by Cleveland or Heer. Without a formal measure, and in need of an estimate, a group of experts was polled, and yielded answers of 3, 4 or 5 perceived levels, with 4 being most common. In the proportional encoding, the step size is minimally a single character, meaning that some small differences for the lowest readerships which are discernable in the smallest treemap boxes are not discernable in the proportional string encoding. Conversely, at the larger sizes, differences between two similar quantities may be difficult to discern in the treemap because areas of different aspect ratios have higher error rates (see Heer), while these differences are rapidly apparent in the proportional string encoding. There are approximately 10 uniquely perceivable sizes in this example.
- *Headline:* is encoded directly as text. However, the treemap encoding results in text that is unreadable visually: if a headline is too small, shorter than 60 characters or does not appear it is considered unreadable, with the same threshold used for *too small* used in all cases. This text could be readable via interaction, however interaction is much slower than a shift in gaze. In the case of the treemap, only 15-20 headlines out of 31 headlines were readable at a target resolution of 640x360 whereas in the other 2 representations all headlines were readable in a display that was only 52% of the area of the original treemap.

Table 1. Number of levels per attribute for each news item visualization variant

Variant	Recency	Readership	Headline
Treemap	3	10-14	15-20
Font Weight	3	4	31
Proportional	3	10	31

These measurements can be repeated for a number of different instantiations of the design or rapid prototype, for example, at more extreme cases. These designs and measures were repeated for two additional variants, one very dense with 56 headlines in a tiny space such that all headlines cannot be displayed regardless of the design alternative (shown in figure 3), and the other variant with 56 headlines displayed sparsely such that there is more than sufficient space to display all headlines. The results are shown in table 2.

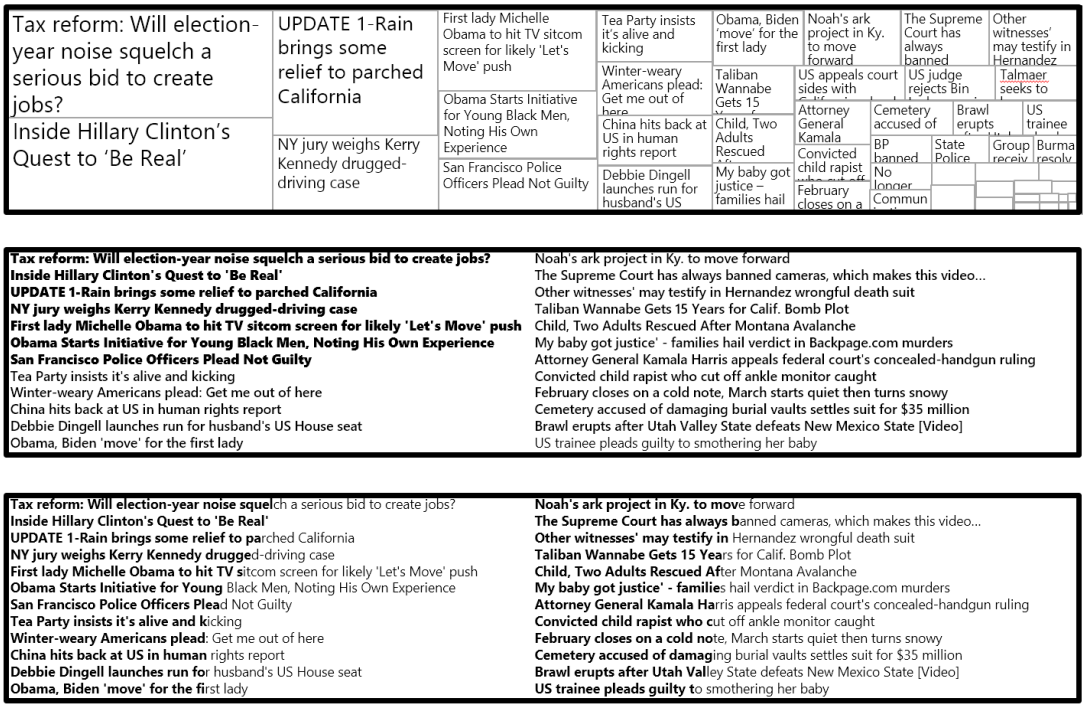

Figure 3. Another example of the news headlines (without background brightness). This example is dense with 56 headlines represented in a small area: in all cases, not all headlines are visible. Top: size indicates readership. Middle: text weight indicates readership. Bottom: proportional length of bold indicates readership.

Table 2. Number of levels per attribute for each news visualization variant for dense and sparse variants

Variant	Recency		Readership		Headline	
	Sparse	Dense	Sparse	Dense	Sparse	Dense
Treemap	3	3	12	12	27	9
Font Weight	3	3	5	3	56	24
Proportional	3	3	12	10	56	24

Computing the average across the scenarios results in the values shown in table 3. Table 3 summarizes the tradeoff between the treemap, which has a high fidelity for readership (i.e. many levels) but loses some headlines, versus the font-weight encoding, which shows a high number of headlines but a low number of readership levels. Note that an average has been used here, but there may be other ways to summarize this data. For example, a plot showing the relationship between the number of levels and relative density could indicate some representations perform well at certain sizes.

Table 3. Average number of levels per attribute across scenarios

Variant	Recency	Readership	Headline
Treemap	3	12.0	17.7
Font Weight	3	4.0	37.0
Proportional	3	10.7	37.0

Initially, the treemap and the font-weight headlines were the only two design alternatives. This tradeoff between treemap versus font-weight headlines was visible in the initial designs and these fidelity metrics facilitated more focused consideration. The result was a design time dissatisfaction with both alternatives and spurred the design of the third alternative (proportional length encoding, detailed in [21]). The proportional length encoding provides a similar high

fidelity for readership, like the treemap, and a similar high fidelity for readership, like the font weight variant; resulting in an overall seemingly superior design which has moved into a more detailed design exploration.

3.3 Overall Lossiness

While the number of levels per each dimension is useful, some means of combining these values together into a single score would be useful to evaluate different design alternatives. Visual attributes can typically be combined together, for example, a bubble plot with bubbles at five different sizes and six different hues can represent 30 different unique combinations of size and color. The various permutations across combined visual attributes is multiplicative, in general, although there are caveat. For example when combining hue and brightness, all hues with a brightness of zero are black reducing the number of permutations. This means in practice, the multiplicative combination of levels per channel represents a maximum potential permutations. The design and the perceivable levels should also take into account these interferences, for example, the hue and brightness combination can be addressed by varying brightness in a narrower range than going all the way to a full black.

A relative comparison of the permutations per design variant then results in a relative lossiness score. This is analogous to the explanation that a graphics card can display 16 million colors (i.e. 256 levels of red x 256 levels of green x 256 levels of blue) even though the card may only support a display size of 1920x1080 - i.e 2 million actual pixels meaning that only a subset of the 16 million colors can be displayed at any one time.

Using this multiplicative approach, the relative lossiness of the three design variants is as follows in table 4. Note that a lower score is more lossy, so the font weight encoding is the most lossy (at 0.7 relative to treemap) and the proportional encoding is least lossy (i.e. more preserved levels of the original data, with a score of 1.87 relative to the treemap.

Table 4. Relative Lossiness of Alternate Headline Representations

Design Variant	Recency	Readership	Headline	**Total Permutations**	**Lossiness Relative to Headline Treemap**
Treemap	3	12.0	17.7	637	1.00
Font Weight	3	4.0	37.0	444	.70
Proportional	3	10.7	37.0	1188	1.87

Based on this technique, the proportional encoding appear to offer the least loss. Font-weight appears to be more lossy than the treemap, although the metric is an "apples and oranges" aggregation: 1) Headlines express complex ideas whereas readership expresses a single quantitative value. Loss of a headline may have a higher weight than the loss of a quantitative value about a headline. 2) Headlines require active cognitive reading to be understood whereas sizes can be understood pre-attentively at-a-glance. While separate fidelity scores can be combined it is still useful to retain the constituent fidelity score per attribute.

4. GENERALIZED FIDELITY AND LOSSINESS

The general approach to calculating fidelity and lossiness is as follows.

4.1 Fidelity Estimation

For each data dimension that is encoded as a visual attribute, the *fidelity* is calculated to determine the number of unique levels perceivable. The number of levels perceivable will the lesser of the number of unique data instances and the maximum number of levels perceivable based on experiments and guidelines.

Example 1: One data attribute from the Titanic dataset (*http://en.wikipedia.org/wiki/Passengers_of_the_RMS_Titanic*) is the class of the passenger, i.e. first, second or third. If this data attribute is encoded as hue, the number of unique levels perceivable is three. Various authors suggest that the maximum number of perceivable hues is eight to ten, but the number of unique instances in this encoding will only be three resulting in a utilization of only three levels for hue.

Example 2: One data attribute from Fisher's Iris dataset (e.g. see *http://en.wikipedia.org/wiki/Iris_flower_data_set*) is sepal length, which has 35 unique values. If this data is encoded as brightness only, the number of unique levels is only four, using Ware's threshold of four levels of brightness. Even though current monitors can show 256 levels of

brightness for gray, a maximum of four will be perceivable. If sepal length were instead encoded as bars compared along a common baseline and sorted, error rates of ± 2.5% suggest that at least 20 levels can be encoded.

To determine the number of levels perceivable, prior work estimating error rates or guidelines are useful. Some examples include Bertin [6]. Cleveland and McGill [10], Heer and Bostock [11], and Ware [2] summarized in Table 5.

Table 5. Visual Attribute Error Rates and Guidelines for Number of Levels

	Estimation Error Rates		**Number Levels Guidelines**	
Visual Attribute	**Cleveland and McGill**	**Heer and Bostock**	**Bertin**	**Ware**
Adjacent positions relative to common scale *(T1 - adjacent bars within a chart)*	2.5%	2.2%	10 per centimeter	
Positions aligned to common scale *(T2 - between stacked bars)*	3.0%	2.5%		
Positions aligned to common scale *(T3 - between clustered bar charts)*	3.5%	2.8%		
Lengths non-aligned *(T4)*	5.0%	4.0%		
Lengths aligned in sequence *(T5)*	6.6%	4.6%		
Angle *(T6 - pie wedge)*		4.5%	4	4
Area *(T7 - circle)*		6.0%	20	
Area *(T8 - rectangles aligned centers)*		5.5%	*Bertin's examples tend to be square*	
Area *(T9 - treemap rectangles)*		6.0%		
Brightness			6	2-4
Hue			8	10
Saturation				3
Texture (multiple attributes: scale, orientation, pattern, contrast)			4-5	4-12
Shape / Glyphs			Infinite *but not preattentive*	32 combination of shape, color, etc

Ideally, the table of visual attributes with error rates and/or guidelines for levels would exist for all visual attributes. Some authors do provide guidelines for specialized attributes, e.g. fonts [22] and useful future task would be to collect, compare and organize more of these values from multiple sources. However, not all visual attributes have experimental error rates nor guidelines. In these cases, ideally, an experiment should be defined to measure the number of levels perceived or the error rate.

In lieu of an experiment, the author has on occasion presented a sample design utilizing the target attribute to an audience of ten or more people with a show of hands to indicate the number of discrete levels perceived. In practice, this approach tends to result in the majority of votes for only one level with majority narrow distribution of votes. The author has repeated this approach with the same example in three different settings with three different audiences and achieved the same results in each case.

Another alternative is to collect user feedback from design review sessions. User feedback on high-quality design mockups can indicate where users believe that a particular encoding may be less effective than the designer believes, such as an inability to easily identify multiple levels of hierarchy in a treemap.

Finally, the design itself with sample data should be visually inspected. The particular combination of visual attributes, the size of the glyphs, the font used, or other interferences may show that the visual attribute has fewer levels of perceivability than anticipated.

4.2 Permutations and Relative Lossiness

Visual attributes can be combined together to show multiple data attributes. This is a multiplicative effect when that the visual attributes do not interfere with one another. For example, hue and shape are independent: three shapes with three different colors yields nine distinct combinations. However, brightness and texture are not independent: texture requires brightness to be visible - therefore using texture together with brightness may reduce the number of levels of brightness

that can be utilized. In general, when using this approach with combinations of visual attributes that are not independent, the impact should be taken into account when computing the number of levels perceivable per visual attribute.

Number of permutations is the multiplication of all the number of levels perceivable into a total permutations. A higher value indicates a greater number of permutations and the potential to carry more information - i.e. less lossiness.

Number of permutations is computed for each design variant. Relative lossiness, is a simple transformation normalizing the number of permutations to a chosen design. A lower relative lossiness score indicates the potential for more information loss. Design alternatives can then be compared relative to the chosen design. Lower lossiness scores indicate lower amount of information retained and higher lossiness scores indicate a higher amount of information is potentially retained.

Note that this lossiness score only measures information loss from the choices of visual attribute encoding. Information can also be lost at subsequent steps in the process of perception. For example, a scatterplot may lose some information due to overplotting in the resulting visualization; and the viewer may have additional information lost in perception, for example, if the viewer has color blindness or if the viewer is unfamiliar with the representation and makes interpretation errors.

5. POST HOC ANALYSIS EXAMPLE

As a comparison, a well-defined design task with a known outcome can be evaluated using this technique to see if a lower lossiness alternative was chosen in practice. This particular design task occurred 8 years ago for a client. This scenario is also interesting because the design alternatives, which required serious implementation effort 8 years ago, can now be prototyped by a wider audience of developers much more quickly and easily using tools such as Protovis, d3.js or WebGL, however, understanding and assessing tradeoffs has not been made easier.

In this example, the user community needed to understand a hierarchy of data, each level with a magnitude and change measurement. For example, the *Consumer Price Index* (CPI), is a hierarchy of prices weighted by the proportion that individuals spend of each item (e.g. gasoline, rent, food), with a percent change in price in each item that can be aggregated up through the hierarchy to a total level. The users are interested in the magnitude and change throughout the levels and the initial starting point was dissatisfaction with a treemap as the hierarchy was not considered visible by the users and the intermediate aggregations were missing (e.g. tomatoes were displayed, but the total for vegetables was not). Figure 4 shows a sample treemap with CPI data.

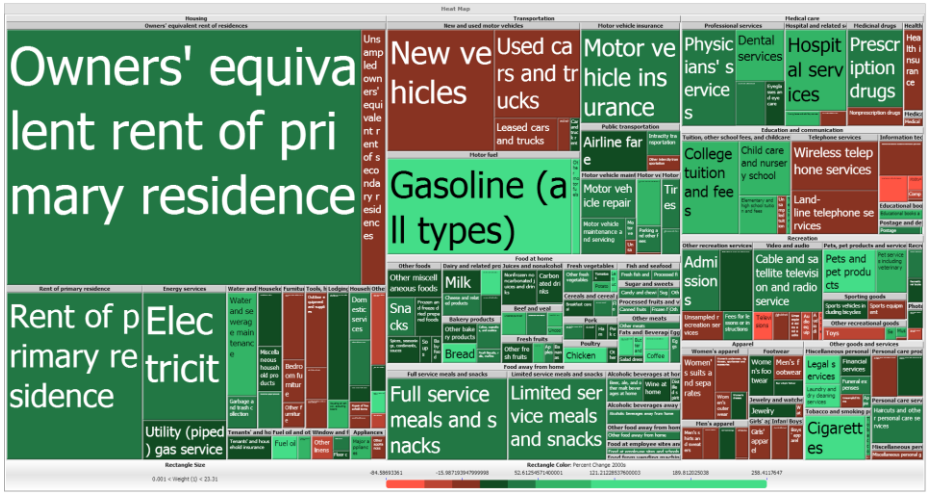

Figure 4. Treemap of US CPI data from www.bls.gov/cpi.

Design alternatives considered included:

- *Treemap*, with size set to magnitude and color set to percent change to prior period. This was the baseline.
- *Voronoi treemap*, with size and color set similarly. Similar to the treemap, users did not consider the hierarchy to be effectively visible.
- *Sunburst chart* [23] (i.e. a multi-level pie chart), with each successive level indicating another level in the hierarchy, with pie wedge size indicating magnitude and color indicating percent change. The sunburst clearly showed the hierarchy, colors and sizes. The only labels were around the perimeter, off the chart, indicating the category corresponding to the first level wedge.
- *2D Grid*. Each cell in the grid belonged to a region (clearly demarked) indicating hierarchy, with cell color indicating percent change and label size indicating magnitude. Labels were simply truncated at the edge of the cell (similar to long label truncation in Excel cells). Given the small sizes, labels indicated only the first three or four letters, except for the top level which contained larger cells and clearly labelled the category. Given the small label size, only three levels of label sizes would be discernable.
- *Grid with 3D bars*. Similar to the grid, with an added thin 3D bar. Bar height provides a greater number of levels than label size, but creates issues with readability of text in 3D and potential occlusion.
- *Org chart*. Each bubble on the org chart was colored by the percent change, each bubble varied in size based on magnitude. The actual design was problematic - at 200 leaf nodes within only 400 pixels only left one or two pixels for the leaves, although some staggering and overlap allowed for 3 different sizes.

The design alternatives are shown in figure 5, using a much simpler dataset - the simple dataset is for diagrammatic purposes and can clearly display short-labels and a shallow hierarchy, whereas the original CPI data contains much longer labels and many more data points.

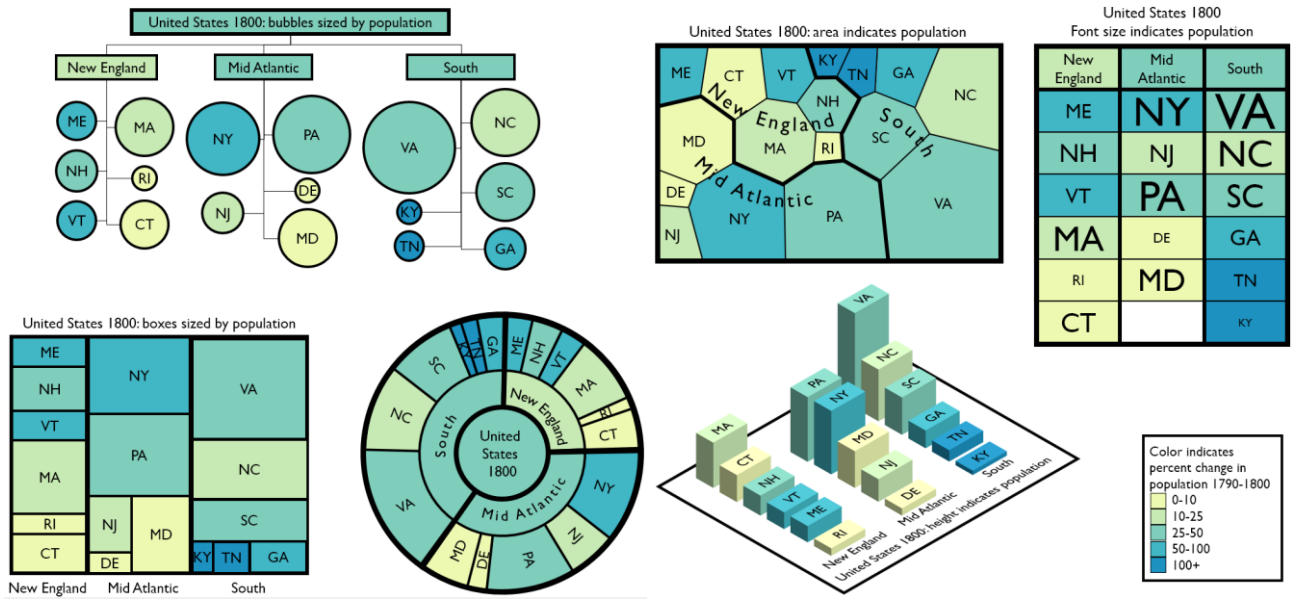

Figure 5. Design variants for depicting a hierarchy. Top row left to right: Org chart, Voronoi treemap, 2D grid. Bottom row: Treemap, Sunburst, 3D bars. Color is consistently applied in all designs.

Levels per each visual attribute, total discrete combinations and relative lossiness (vs. treemap) are shown in table 5. The hierarchy column indicates the number of levels of hierarchy clearly visible. In the case of the treemap, users felt that the hierarchy was not particularly visible, therefore it scored only one level. In the case of the grid and 3D bar, clear amounts of whitespace or boundaries delineated the top level hierarchy, but the approach did not extend well through multiple levels, thus providing only two levels. The sunburst and org chart adequately displayed three levels of hierarchy.

Table 5. Levels per attribute and estimated lossiness of design alternatives.

Design Alternative	Hierarchy	Labels	Size	Color	Total Permutations	Lossiness Relative to Treemap
Treemap	1	25	12	7	2100	1.0
Voronoi Treemap	1	10	12	7	840	0.4
Sunburst	3	10	16	7	3360	1.6
Grid	2	50	3	7	2100	1.0
3D bar	2	10	16	7	2240	1.1
Org Chart	3	25	3	7	1575	0.7

The label column indicates the number of readable labels. The treemap is clearly capable of displaying many labels, as is the org chart (oriented left to right). The grid with truncated labels, however, is difficult to quantify using this approach. To a novice user, most labels truncated to only three characters would be useless, but for a domain expert they could be useful. Clearly some labels will be ambiguous regardless of viewer, e.g. *oth(er meats)* vs. *oth(er foods);* and even domain experts may need more than a three letter cue for obscure categories. It is added here at 50, on the assumption that 50 of these severely truncated labels will be useful to viewer somewhere between novice and expert.

The size column indicates the number of levels of sizes distinguishable based on metrics from [10,11]. Based on [11], angle outperforms area estimation and the number of levels is proportionally higher for sunburst vs. treemap. Similarly, judging lengths with a common scale but not aligned with a baseline performs even better than angle and proportionally should have a value of 20 levels, however, these are 3D bars and the levels will be reduced by perspective effect and potential occlusion and therefore has been reduced to the similar level as sunburst. Note that the 3D bars also have the same labels as the grid, but given the small size, perspective distortion and occlusion, only the top level labels are considered in the table.

The color column has been set uniformly to seven levels, assuming 3 levels of green and 3 levels of red and one neutral color are visible. As all levels are the same, this column essentially has no effect on the total permutations column. The total permutations column is simply the multiplication of the preceding columns to express the total number of discrete combinations, and the final column adjusts these numbers relative to the treemap starting point.

In practice, the sunburst chart was implemented, has been successful and this is the chart that has the best lossiness score in the table above. However, suppose that the user goals and objectives were not known in advance and thus discretionary judgments could be made differently. For example, instead of using levels of hierarchy, suppose the number of nodes displayed was used: as there are fewer nodes with each successive move up the hierarchy, the number of levels of hierarchy shown would in effect be weighted lower. If users are assumed to be expert users, the extremely short labels could be effective - perhaps 100 or even 200? Similarly, if users are assumed to be adept at 3D height perception the number of discrete heights could be close to 20. These adjustments result in values shown in table 6.

Table 6. Alternative calculations for levels per attribute and estimated lossiness of design alternatives.

Design Alternative	Hierarchy	Labels	Size	Color	Total Permutations	Lossiness Relative to Treemap
Treemap	200	25	12	7	420000	1.0
Voronoi Treemap	200	10	12	7	168000	0.4
Sunburst	300	10	16	7	336000	0.8
Grid	210	150	3	7	882000	1.6
3D bar	210	10	20	7	294000	0.7
Org Chart	300	25	3	7	157500	0.4

In this alternate post hoc analysis, the grid performs best. However, any design time decisions regarding encoding must be considered in the higher level context of the domain and task [17,24]. In this scenario, even if the above metrics are generated without the context of the requirements and the requirements are discovered at a later stage, then the grid is ruled out (because of required expertise) and the treemap is ruled out (because of user dissatisfaction) leaving the sunburst as the best remaining alternative. Thus the approach can work, although user goals and tasks and necessary, and best applied when making judgments about fidelity per attributes.

6. GENERAL INFORMATION VISUALIZATION EXAMPLE

Up to this point, fidelity and lossiness has been applied to specific visualizations with specific datasets. Can the approach be applied more generally to generic information visualizations without data? For example, a dataset with three quantitative attributes could be represented as a parallel coordinates chart, a bubble plot, or a set of star glyphs.

In the case of the parallel coordinates chart, each quantitative value can be mapped to a position along an axis. Using Bertin's rate of one millimeter on a typical laptop screen this may be on the order 150 unique positions for each attribute. Total permutations are 150 x 150 x 150 = 3,375,000.

In the case of the bubble plot, assuming the same screen, the X and Y axis have similar number of unique positions. However the relative sizes of the bubbles have an error rate of 6% and Bertin suggests 20 sizes. Total permutations for the bubble plot are 150 x 150 x 20 = 450,000.

In the case of star glyphs, each glyph has three axes at different orientations, with the length of each axis indicating a variable. If each glyph is placed separately so that glyphs do not overlap, the number of glyphs determines the size of each glyph, which in turn impacts the number of the levels that can be perceived. Assuming 500 items, each glyph has approximately 1/20th the width and height available, resulting in only 8 or so levels per data attribute. The permutations are 8 x 8 x 8 = 512.

A comparison of these values shows that the parallel coordinates approach has the least lossiness, followed by the bubbleplot with star glyphs in a distant third place. In this generalized example, the parallel coordinates approach is superior.

What is missing in this comparison is any notion of the data or the tasks required. Parallel coordinate displays the data with a high degree of fidelity per attribute, but visual separation between elements can be difficult (which is much easier both the scatterplot and star glyph displays). Some types of patterns can be easier to see in a scatterplot or star glyph than a parallel coordinate display. The star glyphs, with clear spatial separation, can more readily support labelling than either the parallel coordinate display or the bubbleplot. If a designer starts instead with a task, such as identifying individual outliers, the parallel coordinate display may not even be included in the design space as it may not be relevant to the task.

This example serves to show the approach here is a design time tool for evaluating comparable visualizations for a target task and not an approach for ranking visualization techniques. Applying this technique without task consideration would be a failure in the data and task abstraction step in Munzner's four level nested design model [17].

7. LIMITATIONS AND FUTURE WORK

The measurement of fidelity and relative lossiness can be useful, particularly when evaluating design alternatives for accuracy (e.g. [10,11]). Many more visual attributes are feasible and it would be useful to extend accuracy research across a wider range of visual attributes and/or have methods for estimating the number of levels supported by a visual attribute in a particular context. As a proxy, as can be seen here, techniques for estimating the number of levels for fidelity include measured values in previous accuracy studies; polling experts or users; or visual inspection of a design that contains data. Some of these values may need to be adjusted based on well-defined thresholds, such as minimum legible screen sizes for fonts; or estimated, such as 3D occlusion reduces readability of some labels.

Understanding the higher level domain problem, including the users goals and tasks, is required to use this approach: this information helps frame judgments that are based on user needs and capabilities, such as ability to read truncated labels or expertise with 3D interfaces. The example provided in the post-hoc analysis provides some indication that there may be some robustness to the approach although the context of user needs and capabilities is required.

No aspect of the user task nor the relative importance of the particular data attribute to the task is directly captured in the metrics of fidelity and relative lossiness. A less important data attribute to the task does not require a dimension with high fidelity and more lossiness is acceptable. Thus, the components of relative lossiness could be to be weighted to match the task.

One particular shortcoming is that this approach only records the information discernable in the target display, not the speed at which the representation can be comprehended. Visual attributes that can be pre-attentively perceived, e.g. length, area, hue, brightness, etc, are mixed together with attributes that require active attention (i.e. text labels). The best

lossiness score would be attainable with a data table resulting in no lossiness, but also would have no information visualization properties, e.g. no information would pop-out, no patterns would be discernable at-a-glance. Similarly, glyphs based on complex shapes can have a high number of levels but also have difficulties with perception: Bertin [6] warns that shape can have infinite number of levels and it is tempting to use it but provides examples where patterns are not visible in fields of complex icons. Amar [15] also warns against *representational primacy* over *analytic primacy* and the ability to perceive patterns depends upon the visual system detecting pattern across pre-attentive visual attributes. Therefore, further extending this approach with the relative speed of perception of different visual attributes would be ideal. However, the simple visual attribute rankings (e.g. [6,7,8,9]) do not clearly identify the reason for the ranking and some high ranking items are high because of accuracy of estimation not necessarily speed of perception (e.g. [7]). Attribute rankings from perceptual psychology could be used instead, for example Wolfe and Horowitz [25] provide a table indicating visual attributes that might guide attention in visual search. Furthermore, all visual channels are weighted equally when creating an overall lossiness score. Weighting could better adapt the model for attributes such as labels or icons which provide a very high number of categorical levels (i.e. each unique label) at a cost of active reading.

The essence of an intuitive display or aesthetic appeal is not captured in fidelity and lossiness metrics. Some other metrics and clustering techniques do attempt to improve the intuitiveness and improve visceral appeal of the result, e.g. [25,26,27]. One visualization expert reviewer agreed with all the logic for measuring the best performing design alternative in figure 2, but commented that the treemap of headlines was still a more viscerally engaging representation that the other alternatives.

In summary, this section suggests that design-time evaluation is multi-faceted and that metrics should include some combination of 1) fidelity/lossiness; 2) speed of perception; 3) intuitive displays and aesthetics; and that these should be considered in the context the user goals and tasks. These metrics could be readily available at design time with appropriate research to quantify levels and performance per visual attribute.

8. CONCLUSIONS

These novel metrics for measuring fidelity and lossiness can be a quick design-time calculation to consider how effective a particular set of visual encodings will be. Additional expert feedback and user evaluation will still be required, but the approach here can be a fast means to quickly pruning the design space to discard highly ineffective configurations or spur visualization designers to more imaginative, higher fidelity encodings.

REFERENCES

[1] Chen, M., and Floridi, L. "An analysis of information in visualization." *Synthese*, 2013.

[2] Ware, C. [Information Visaulization: Perception for Design], Morgan Kaufmann, Waltham, MA, 2013.

[3] Forsell, C. and Johansson, J. "An heuristic set for evaluation in information visualization," *Proceedings of the International Conference on Advanced Visual Interfaces*. 199–206. ACM, (2010).

[4] Freitas, C. M. D. S., Luzzardi, P. R. G., Cava, R. A., Winckler, M. A. A., Pimenta, M. S., and Nedel, L. P. "Evaluating usability of information visualization techniques." *Proc. 5th Symposium on Human Factors in Computer Systems (IHC) 2002*, 40–51, (2002).

[5] MacKinlay, J. "Automating the design of graphical presentations of relational information." *ACM Transactions on Graphics (TOG)*, 5(2), 110-141, (1986).

[6] Bertin, J. [Semiologie Graphique], Gauthier-Villars, Paris, (1967).

[7] Cleveland, W. [Elements of Graphing Data], Hobart Press, Summit, NJ. (1985).

[8] Wilkinson, L. [The Grammar of Graphics], Springer, (1999).

[9] Mazza, R. [Introduction to Information Visualization], Springer, (2009).

[10] Cleveland, W., and McGill, R. "Graphical perception: Theory, experimentation, and application to the development of graphical methods." *Journal of the American Statistical Association*, 79(387), 531-554. (1984).

[11] Heer, J. and Bostock, M.. "Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design." *ACM Human Factors in Computing Systems (CHI)*, 203–212, (2010).

[12] Beshers, Clifford, and Steven Feiner. "Autovisual: Rule-based design of interactive multivariate visualizations." *Computer Graphics and Applications, IEEE* 13.4 (1993).

[13] Senay, H., and Ignatius, E.. "A knowledge-based system for visualization design." *Computer Graphics and Applications, IEEE* 14.6 (1994).
[14] Senay, H., and Ignatius, E. [Rules and Principles of Scientific Data Visualization]. Institute for Information Science and Technology, Department of Electrical Engineering and Computer Science, School of Engineering and Applied Science, George Washington University, 1990.
[15] Amar, R., and Stasko, J. A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations", IEEE Symposium on Information Visualization. (2004).
[16] Lam, H., Bertini, E., Isenberg, P., Plaisant, C., and Carpendale, S.: "Empirical studies in information visualization: Seven scenarios." *Visualization and Computer Graphics, IEEE Transactions on* 18, no. 9. (2012).
[17] Munzner, T. "A nested model for visualization design and validation." *Visualization and Computer Graphics, IEEE Transactions on,* 15(6), 921-928. (2009).
[18] Bertini, E., Tatu, A., and Keim, D. "Quality metrics in high-dimensional data visualization: an overview and systematization." *Visualization and Computer Graphics, IEEE Transactions on* 17, no. 12 (2011): 2203-2212.
[19] Brath, R. "Metrics for effective information visualization." *Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis' 97),* (1997).
[20] Weskamp M. "Projects: Newsmap", 2004. URL: http://marumushi.com/projects/newsmap. 2, 7 (03/03/2014)
[21] Brath, R. and Banissi, E., "Using Font Attributes in Knowledge Maps and Information Retrieval", *Proceedings of Knowledge Maps and Information Retrieval (KMIR) at Digital Libraries 2014,* CEUR, (2014).
[22] Brath, R. and Banissi, E. "The Design Space of Typeface", *Proceedings of the 2014 IEEE Symposium on Information Visualization (VisWeek 2014),* (2014).
[23] Stasko, J., Catrambone, R., Guzdial, M., and McDonald, K. "An evaluation of space-filling information visualizations for depicting hierarchical structures." *International Journal of Human-Computer Studies*, 53(5). (2000).
[24] Meyer, M., Sedlmair, M., and Munzner, T. "The four-level nested model revisited: blocks and guidelines." *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors-Novel Evaluation Methods for Visualization.* ACM, 2012.
[25] Wolfe, J. and Horowitz, T.. "What attributes guide the deployment of visual attention and how do they do it?" *Nature Reviews, Neuroscience*, 5(6), 495-501. (2004).
[26] Bertini, E., and Santucci, G. "Quality metrics for 2d scatterplot graphics: automatically reducing visual clutter." *Smart Graphics* (77-89). Springer. (2004).
[27] Peng, W., Ward, M. O. and Rundensteiner, E. "Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering." http://digitalcommons.wpi.edu/computerscience-pubs/71. (2004).
[28] Wilkinson, L., Anand, A. and Grossman, R. L. "Graph-Theoretic Scagnostics." *INFOVIS*. Vol. 5. (2005).